DOCUMENT RESUME

ED 081 781 TM 003 141

AUTHOR	Katz, Martin
TITLE	Selecting an Achievement Test: Principles and Procedures.
INSTITUTION	Educational Testing Service, Princeton, N.J.
PUB DATE	73
NOTE	12p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Achievement Tests; Educational Testing; *Norms; Standardized Tests; *Test Reliability; Test Results; Tests; *Test Selection; *Test Validity

ABSTRACT

The recommended procedure for achievement test selection is presented. Following a review of some fundamental definitions of statistical properties of achievement tests (validity, reliability, and norms), major considerations to be applied in selecting achievement tests are discussed. These considerations are: (1) the characteristics of the school population, (2) the content and objectives of the curriculum, (3) the purposes of testing, and (4) the use of test scores. An appendix provided a list of U.S. publishers of standardized tests who issue test catalogs. (DB)

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

SELECTING AN ACHIEVEMENT TEST:

Principles and Procedures

by Martin Katz

EDUCATIONAL TESTING SERVICE, PRINCETON, NEW JERSEY

ED 081781

TM 003 141

Copyright © 1958, 1961, 1973 by Educational Testing Service. All rights reserved.

School people who are responsible for selecting standardized achievement tests often ask: What is the best test to use for such-and-such a subject or grade? or, What is the best achievement testing program for a school system or college? Good answers to these questions are essential to the effectiveness of a testing program. Many problems in test use and test interpretation have their origin in test selection.

However, good answers cannot be drawn from stock. Schools differ so that no one test or one program is equally suitable for all. Certainly, the measurement expert who does not know the individual school should not, in good conscience, try to "fit it in the dark." Fortunately, a large quantity and great variety of standardized achievement tests are available, permitting each school to tailor a program to its own needs and characteristics. But how can educators who are not measurement experts be helped to choose from the vast array of materials?

Shifting the emphasis from "what" to "how" offers a solution. While no tests can be prescribed as "best" for all schools, there are fundamental principles and procedures for test selection that can be recommended to all. The particular program of tests chosen may be different for each school, applicable to no other; but the process of choosing wisely seems to contain universal elements, applicable by many schools.

These elements include careful inquiry into the following topics: I, the nature of the pupil population; II, the content and objectives of the curriculum; III, the purposes in testing; IV, the ways in which test scores may be used to accomplish those purposes; V, the readiness, willingness, and ability of the school staff to administer tests and interpret scores; VI, the amount of time and money available for testing.

Questions and answers about school staff and time and money (items V and VI) will not be elaborated here. Such circumstances can be quite readily described, and their bearing on test selection quickly understood. It is the other four points—pupil population, curricular objectives and content, testing purposes, and prospective uses of scores—that usually need more specific explanation. Merely to admonish that they "should be taken into account" does not seem sufficiently helpful. Therefore, this pamphlet will try to explain *how* they "should be taken into account." It will discuss in some depth and detail how to analyze these four major considerations as they relate to sound measurement principles in the selection of standardized achievement tests.

For to many school people "testing achievement" seems in itself like an adequate statement of purpose and use, and the designation of grade level and subject area seems like an adequate description of population and content. They assume that if a test measures achievement it measures achievement, and any application or use can be made of the results with equal relevance. And of course by this naïve reasoning the test which "best" measures achievement is the best one to choose.

Those who have had some exposure to measurement lingo may specify this "best" achievement test they are seeking as the one with the "best" validity, the "best" reliability, the "best" norms. However, their search on these grounds is likely to be confusing and futile because the question—What is *the best* achievement test?—is probably the wrong question. It is more profitable to ask, What test can be most effectively used in a certain way for a designated purpose in a specific school setting? This kind of inquiry shapes more distinctly the questions that should be directed to validity, reliability, and norms. Since these statistical properties of tests must figure in our later discussion of pupil population, curriculum content and objectives, purposes in testing, and uses of test scores, let us review briefly here a few fundamental definitions and concepts that can serve as guide lines and warnings.

Definitions and concepts: statistical properties of achievement tests

Validity is often defined as the extent to which a test measures what it is supposed to measure or does the job it is given to do; *reliability* as the extent to which a test is consistent in measuring what it measures; *norms* as the statistics describing the test performance of a specific group which can be used for comparison.

These test characteristics are frequently reported in quantitative terms. For example, a per cent agreement or correlation coefficient (indicating degree of relationship) may figure prominently in validity studies; a correlation coefficient almost always appears as an index of reliability; and a description of norms generally stresses the number of people tested to establish a comparison group.

Sometimes prospective test users, upon being advised to "take into account" the validity, reliability, and norms of tests, are tempted to make a direct comparison of these indices and numbers from one test to another. Some will assume, for example, that if in a certain study or series of studies, Test A had a higher correlation with teachers' marks than Test B, Test A is generally more valid than Test B; that a test which is reported to have a "reliability coefficient" of .93 is more reliable than a test which is reported to have a "reliability coefficient" of .88; that norms based on test performance of 100,000 pupils are "better"—more likely to be a "national sample"—than norms based on test performance of 25,000 pupils. Then, if the test they have selected by these erroneous assumptions as the "best" fails to work out effectively for them, they are likely to view with suspicion the publisher's claims about validity, reliability, and norms. They wonder whether he has been guilty of misrepresentation or statistical skulduggery.

Test publishers who try to follow the *Technical Recommendations for Achievement Tests*[1] in their manuals and

[1]Published by Committee on Test Standards, American Educational Research Association (AERA) and Committee on Test Standards, National Council on Measurements Used in Education (NCMUE), January 1955.

FILMED FROM BEST AVAILABLE COPY

other publications do much to dispel such suspicions. The *Technical Recommendations*, speaking for the educational research and measurement professions, furnishes standards of approved practice for the guidance of both users and producers of tests. These standards specify what essential information on validity, reliability, and norms (as well as interpretation, administration, and scales) should usually accompany a test to enable users to evaluate it.

Validity

The *Recommendations* will alert the test user to the type of validity and to the kinds of substantiating data he should look for. Four types of validity are defined.

1. "*Content validity* is concerned with the sampling of a specified universe of content."

Sometimes, content validity is completely self-evident. For example, suppose we want to measure pupils' recall of the alphabet, or the names of the 50 states, or Latin verbs that take the dative, or the "multiplication table," or the correct spelling of a certain list of words. We can present the complete task itself to the pupils. In such cases, the entire test is identical with—not merely representative of—the knowledge or skill to be measured.

More often, the knowledge or skill that we are interested in measuring cannot be presented *in toto*. We want to know more about the pupils' reading, social studies, Latin, arithmetic, or spelling achievement than is usually encompassed in such limited tasks. If, for a spelling test, we were to use all the words in an unabridged dictionary (which, for illustrative purposes, we can take to comprehend the entire universe of English words), we should have an unconfortably long list. Furthermore, this long list is probably not the universe that we are interested in. For example, if we want to test the spelling of 8th-graders, we are not likely to be concerned with words like "syzygy" or "xyster," rarely encountered or used by 8th-graders. Nor are we likely to be concerned with words like "cat" or "boy," which we may prefer to assume all of our 8th-graders can spell. Perhaps we should rather focus on a universe (or solar system) consisting of all words misspelled in the written compositions of a certain large number of 8th-graders, or the words used by them with a certain degree of frequency, or words contained in a certain array of reading materials, or some combination of these. Such a procedure still gives us a long list. But at least it is a list which we can define in specific terms. Then, we can take a *sample* of that list, selected so as to truly represent it. In other words, pupils should spell correctly the same proportion of words from the sample as they would from the complete list. If so, the sample list, or test, has high *content* validity.

2. But we cannot expect use of the sample list to tell us any *more* about a pupil's spelling ability than use of the original list would have told us. Therefore we shall probably have questions concerning its *concurrent validity*, which "is concerned with the relation of test scores to an accepted contemporary criterion of performance on the variable which the test is intended to measure." For example, how similar is the 8th-graders' performance on the test-task (say, writing words from dictation, or judging whether a printed word is correctly or incorrectly spelled, etc.) to the kind of spelling skill they show in their own writing? Of course, in writing a composition, some pupils will look up spellings they are not sure of; others will take a guess; others will limit their written vocabulary to words they feel sure they can spell. Or pupils may even have different spelling habits in different types of composition. Nevertheless we shall probably expect to find some relationship between spelling test scores and pupils' spelling in their own compositions. Data demonstrating such relationships will be evidence of concurrent validity.

3. Research studies showing a relationship between performance on the spelling test and performance in some future activity will be evidence of its *predictive validity*—which "is concerned with the relation of test scores to measures on a criterion based on performance at some later time." Thus, as an example of predictive validity, there may be evidence that scores on this spelling test can help us to differentiate between pupils who will succeed and pupils who will fail in a stenography course. Time—relating test performance to a present or to a future criterion—is the chief difference between concurrent and predictive validity.

4. When direct criteria are not available for evidence of concurrent or predictive validity, and the universe of content can not be adequately specified, it is necessary to rely more heavily on *construct validity*—logical inferences which can be drawn from indirect evidence that the test measures what it is claimed to measure. We may ask, what hypotheses about the nature of spelling ability does this test reflect? Do test performances of pupils tend to confirm these hypotheses? In what ways do high scorers differ from low scorers? How, for example, are scores related to methods of instruction in reading? What is the factorial composition of this test? Thus, evidence of construct validity is likely to be both more varied and more theoretical than evidence of the other types of valid

Evidence of each type of validity answers a somewhat different question about the test: **content**—what does the test represent and how truly representative is it? **concurrent**—with what other present observations of pupils' knowledge, understanding, or skill do scores on this test agree and how closely do they agree? **predictive**—what kinds of future performance can be predicted by scores on this test and how accurately can they be predicted? **construct**—what else do we know about the content, relationships, and rationale of this test?

The sophisticated test-user realizes, then, that validity is not some general, pervasive characteristic that a test "has." Rather, the validity of any test is likely to be specific to the job he wants that test to do. Whether he requires evidence of content validity, concurrent validity, predictive validity, or construct validity depends on his purpose. Whether the evidence is convincing depends on its relevance to the setting in which he expects to use the test. An approach to judging the validity of a test, then, will be implicit in our discussion of pupil population, curriculum content, purpose in testing, and use of scores.

Reliability

The test-user may expect to find himself concerned with one or more of three kinds of reliability coefficients:

"PERMISSION TO REPRODUCE THIS COPYRIGHTED MATERIAL HAS BEEN GRANTED BY

Dorothy Urban

TO ERIC AND ORGANIZATIONS OPERATING UNDER AGREEMENTS WITH THE NATIONAL INSTITUTE OF EDUCATION. FURTHER REPRODUCTION OUTSIDE THE ERIC SYSTEM RE-

1. *Coefficient of internal consistency* shows consistency of performance on different parts or items of the test taken at a single sitting. This coefficient is usually computed by "split-half" correlation (e.g., Spearman-Brown formula), or by item analysis (e.g., a Kuder-Richardson formula).

2. *Coefficient of equivalence* shows consistency of performance on different forms of a test. It is based on the correlation between performances on the different forms administered at nearly the same time.

3. *Coefficient of stability* shows consistency of performance on a test (or on equivalent forms) over a period of time. It is based on the correlation between performances on initial test and retest after a distinct interval.

It is clear that each of these coefficients gives a somewhat different view of reliability, including some different sources of measurement error. Whether we should look to one or another of these coefficients in forming judgments about the reliability of a test depends on how we expect to use the test.

What else must we know about these coefficients to judge whether a test is "sufficiently reliable" for our needs? Let us plant a few seeds of statistical lore concerning reliability now and harvest them later in our discussion of pupil population, curriculum content, purpose in testing, or prospective use of test scores:

1. For speeded tests a coefficient of internal consistency can be quite misleading. In considering any reliability coefficient for a speeded test, we must be sure that speed is one of the important elements we want to measure. Otherwise, the index of reliability will be spuriously high. [content]

2. Other things—such as item quality—being equal, the size of a reliability coefficient can be expected to vary with the number of items in a test. But there is a law of diminishing returns—beyond a certain point, increases in the length of a test may produce relatively small increments in a reliability coefficient. Thus, although "high" reliability is desirable, we must ask ourselves how much additional testing time is a given increase in reliability worth? [purpose]

3. Reliability coefficients can be expected to increase with the performance range or "spread" of the sample of people tested. Conversely, the more homogeneous the group in respect to test scores—i.e., the more restricted their range or spread of achievement—the more likely the rank order of their scores is to change from one part of a test, one form of a test, or one taking of a test to another, and the lower the reliability coefficient will be. [population]

Thus, a publisher may report a high reliability coefficient based on a sample that represents a wide range of ability. But in our use of the test we may be concerned with the scores of individual pupils in a single class representing a much narrower range of ability. Clearly, we cannot expect the test to be so reliable for this group as for the publisher's sample. Therefore, some publishers have adopted the practice of reporting "within class within school" reliability coefficients. These may not seem so impressive in absolute magnitude but are much more realistic for such test usage. [purpose]

4. Test performance and consequent reliability of test scores are also likely to vary with other characteristics of the group tested. For example, a test that is very difficult for a group will result in much guessing, many chance scores, and a low reliability coefficient; whereas a much higher reliability coefficient might be derived from the same test taken by a group which did not find it so difficult. (Thus—taking items 3 and 4 together—when we want to compare the reliability of different tests, we should not simply stack the reliability coefficient reported for one test against the reliability coefficient reported for another. In order to evaluate a reliability coefficient, we should know something about the "reliability sample" on which it is based. Then we should ask, How does the group we want to test compare with the reliability sample for each achievement test we are considering?) [population]

5. An index of reliability which is not governed by the nature of the reliability sample is the Standard Error of Measurement. We must bear in mind, however, that the Standard Error of Measurement is expressed in the score units of each test. Therefore, we can best use the Standard Error of Measurement not in direct comparisons between different tests but in judgments about reliability of various scores and score differences on any test. For this purpose, we may do well to note whether the recommended system of reporting scores for a test utilizes the Standard Error of Measurement. Does it thereby help us to avoid making over-fine differentiations between scores of individuals or groups? Also, if the test is a "wide-range" one, we may note whether score reporting allows for variation in size of the Standard Error of Measurement at different score levels. [use]

6. For any test, average scores of groups are more reliable than scores of individuals. Therefore, a test which may not seem sufficiently reliable for evaluating a score difference of a certain size between two pupils may have satisfactory reliability for evaluating an average score difference of the same size between two schools. Or, to put it another way, a test permits finer distinctions to be made among mean achievement scores of groups than among achievement scores of individual pupils. [purpose and use]

7. The nature of the decision to be based on test scores—the degree to which it is tentative or irreversible, coarse or precise—may influence the order of reliability (or validity) required. For example, a final placement or selection test should be more reliable than a preliminary screening test. [purpose]

Norms

If we want to use the publisher's norms provided for any test, our first question should be: What is this norms groups supposed to represent? For instance, is it supposed to be a sample of all 4th-graders in the nation? all the public high school students of a given state who have completed a course in American history? all entering female freshmen in New England teachers' colleges? all male independent secondary school students who have completed two years of algebra?

Norms provide a specific *comparison group*, in relation to which test performance may be meaningfully described. Therefore, whatever the norms group is intended to represent, its characteristics should be clearly defined so that we can judge whether the norms given in the manual will

be useful in interpreting scores for the population we want to test. (Incidentally, one of these characteristics should be the time of testing. The distribution of scores for 4th-graders tested in the spring may not look like the distribution of scores for 4th-graders tested in the fall.)

As for the adequacy of the sample, we should probably be concerned more with the basis for sampling than with the total number sampled. For instance, although there is probably no such thing as a truly "national" sample, national norms can be more closely approximated by testing a few pupils at each of a large number of schools (classified according to geographic region, community size, and other specified variables) than by testing a much larger total number of pupils from relatively few schools. Of course, for many purposes local or other specified norms will be much more useful than "national" norms. Local norms will generally have to be developed by the test user himself. (Some test manuals offer specific directions for compiling local norms.)

Armed with these concepts concerning validity, reliability, and norms, we can avoid making unwarranted comparisons among tests. And, thanks to the influence of the *Technical Recommendations*, more and more test publishers are furnishing the essential information about a test to enable the "qualified user to make sound judgments regarding the usefulness of tests and the interpretation of test data." However, it is still up to the test users to relate the published information to: I, the characteristics of their own school population; II, the content and objectives of their own curriculum; III, the purposes of their own staff in testing; IV, the use of test scores in their own school setting.

Let us suppose that we are members of a test selection committee. What should we know about these major considerations, and how should we apply them to selecting achievement tests?

Although these four topics will serve as separate headings, it is clear that they overlap and intermingle to such an extent that no one of them can be treated without reference to the others.

I. School Population

We can take it for granted that such characteristics of the group to be tested as age, grade, subjects studied (or some combination of these) are limiting factors in selection of achievement tests. We may also wonder whether some tests can be expected to have a high enough "ceiling" for our group or a low enough "floor"; whether we can use published norms; whether we can make reliable discriminations between our pupils' scores; whether the scores are likely to give valid predictions of future achievement.

Since one of the purposes of achievement testing often is to evaluate the educational status of a group of students, it may seem like prejudging or circular reasoning to use characteristics bearing on that status as a point of departure in test selection. However, if we can estimate in advance the general level of our group, we may be able to select a test that is pitched (or "peaked") at that level. For example, there is research evidence that such school characteristics as average class size, per cent of graduates going to college, size of community, geographical region, presence of community library, per cent of fathers who are high school graduates, and amount of financial support are related to certain English, social studies, and science achievement test scores of secondary school pupils. A number of these characteristics are, of course, closely interrelated and are also closely related to measures of aptitude. The first two, however, seem to possess a distinct relationship to achievement test performance even when aptitude is held constant. Thus we can, with some notion of how our school stands with respect to other secondary schools on aptitude test scores, average class size, and per cent of graduates going to college, establish some level of expectation for our school's status on most achievement tests. This expectation can help us to estimate how difficult a given test may be for our group: e.g., a science test that is too hard for one school's seventh-graders might be just right for another school's sixth graders. It can also help us to determine what kind of norms we may seek: e.g., should we choose a test for which there are independent secondary school norms? It can also help us to judge whether to expect reliability comparable to that reported in the manual: i.e., how does our group compare with the reliability sample? It can also help us to weigh the evidence reported for predictive effectiveness of a test: e.g., does our group resemble closely enough the group used for a prediction study so that we may reasonably take the findings to be relevant?

In short, we can describe our own population in terms of the variables that test publishers increasingly use to describe their norms groups, reliability samples, and cases used in validity studies. We can then judge whether the difficulty level of a test is appropriate for our population and can also make an informed preliminary estimate of the applicability of published norms, reliability statistics, and reported validity data.

II. Curriculum Content and Objectives

Standardized achievement tests offer us a systematic way of observing pupils' responses to a call for use of some knowledge, skills, or understandings. Very well then, *what* knowledge, skills, and understandings are we interested in observing, or measuring? Before deciding what we want to test (especially if our purpose is evaluation—and most achievement testing is concerned with some kind of evaluation), we must have a clear identification of what we want to teach. We can immediately notice obvious differences in the areas which tests or test batteries purport to cover—for example, Listening Comprehension, Science, Writing, Social Studies, and Work Study Skills are titles that may appear in some achievement batteries but not in others. On the other hand, Reading Comprehension and Arithmetic or Mathematics (or what appear to be synonyms of these titles) can be found in almost all.

In deciding which areas we wish to test, we shall probably take into account the importance of these areas at various levels in our curriculum. Then, from Buros' invaluable *Mental Measurements Yearbooks*[2] from various annotated bibliographies of tests (such as those issued for

[2]Buros, Oscar K. (Ed.) *The 1938 mental measurements yearbook*, New Brunswick, N.J.: Rutgers University Press, 1938. *The 1940 mental measurements yearbook*, Highland Park, N.J.: The Mental Measurements Yearbook, 1941. *The third mental measurements yearbook*, New Brunswick, N.J.: Rutgers University Press, 1949. *The fourth mental measurements yearbook* (1953). *The fifth mental measurements yearbook* (1959). *The sixth mental measurements yearbook* (1965). Highland Park N.J.: The Gryphon Press.

state testing programs[3], and from test publishers' catalogs,[4] we can make a rather long preliminary list of tests that seem to bear on each of these levels and areas.

Of course, we shall immediately want to extend this kind of comparison beyond the obvious differences in test titles. For tests with identical titles may often tap skills and understandings which are perceptibly different. Some of these differences are very easily spotted. We can see at a glance whether a certain arithmetic computation test includes long division or fractions or decimals. It is a little more difficult to judge at a glance precisely what different skills, concepts, and understandings are included in a group of tests all labelled Reading Comprehension, or Science, or Social Studies, or Arithmetic Reasoning. For example, to what extent do the exercises in a reading comprehension test require the pupil to identify or repeat in some form something specifically stated in a passage (i.e., what the author said)? To what extent do they require him to make interpretations and draw inferences (e.g., what the author's purpose was)? To what extent does a science or a social studies test emphasize recall of specific facts (definitions, names, dates—e.g., the main products or location of a country or state)? To what extent does it call for the application of knowledge in solving problems (concepts and relationships—e.g., the effects of geographical environment on institutions and ways of living)?

Such differences between tests are likely to be fundamental. They seem to pertain to different kinds of knowledge, different levels of understanding, and different complexities of skill. Which kind of test is better? It depends on what kind of knowledge or understanding or skill or other outcomes of teaching we want to measure at this testing.

Perhaps it is important to make explicit here a little declaration of independence: no test tells us what to teach; instead, our own teaching goals tell us what we want to test. Nor do test publishers themselves have any desire to shape our curricula. Instead, they regard our curricula as a challenge to them to provide suitable tests. Thus, the differences among standardized achievement tests for each level and area, although perhaps complicating the problem of selection, enlarge our freedom to select tests that match our goals. There is no question of which is to be master. No matter what test we use, we will not "teach to the test"; we will continue to teach what we (and the community we represent) deem most important. And in selecting an achievement test, we shall generally look for one that seems to measure this content. The variety we note—and should welcome—among achievement tests tends to reflect some of the wide-ranging variations among school curricula and objectives.

So it seems clear that, before we select a test, each test has already—in a sense—been "selected" by its writers. From a vast conglomeration of texts, curricula, syllabi, practices, opinions, and philosophies in each area and at each level of American education, they have chosen to focus on certain instructional outcomes. Therefore, it is often wise to read first the publisher's or writer's own statement of what the test is designed to measure. If this, on the face of it, seems to be in line with the content and objectives of our curriculum, we shall probably want to consider that test further. Then, in Buros' *Mental Measurements Yearbooks*, we can read a description and evaluation of the test (unless it has been published since the latest *Yearbook*) by an "expert" reviewer. The review may alert us to characteristics of the test (or the accompanying research) which we might not otherwise have noticed. Our next step will be to examine a sample copy or specimen set of each test or battery.

At this point it will be very helpful for us to have written down a clear-cut statement of our curriculum objectives and content in each area that we want to test. For the sake of test selection we shall confine this statement to the cognitive domain—that is, pupils' knowledge and understanding. (This limitation does not minimize the importance of the affective learnings—e.g., pupils' attitudes and values; it merely recognizes that the affective domain has not as yet been satisfactorily covered by standardized achievement tests.)

There is an excellent classification scheme that can help us to list our teaching goals in terms of pupils' knowledge, understandings, and skills for any subject area or grade level. This *Taxonomy of Educational Objectives*[5] progresses from the simplest forms of knowledge, such as recall of words and facts, through highly complex intellectual abilities and skills involving comprehension, application, analysis, synthesis, and evaluation. One of the virtues of this comprehensive system is that it lends itself equally well to classification of teaching objectives and to classification of achievement test items.

If we prefer to devise our own classification system, the *Taxonomy* will at least provide us with helpful hints enabling us to be more confident that our omissions are intentional.

We may wish to weigh or rank each of our objectives according to the importance we attach to it. Then we can match the items in any given test with our weighted or ranked list of objectives. We do this by taking the test ourselves and judging what kind of knowledge, understanding, or skill each item seems to evoke. (Some publishers help us by providing item classifications.) We then write the item number opposite the appropriate teaching goal in our list of objectives. If an item does not seem related to any of our teaching goals, we can mark that item number at the bottom of the page opposite the label "Irrelevant." Then, for each test we can examine the quantity of items clustered for each objective. Does this quantity correspond at least roughly to the importance we attach to that objective? In other words, is there a relatively high proportion of items for our most important goals? A low proportion for the goals we deem least important? A low proportion of "Irrelevant" items? If so, we shall want to give this test further consideration. If, on the other hand, the items do not seem to fit the content and objectives of our curriculum, we can assume that this standardized achievement test is likely to have very limited usefulness *for us*.

Although we are probably not yet ready to make a final selection of a test, this initial screening has at least narrowed the field. Some of the remaining tests may, upon

[3]For example, Hastings, J. T. *Test scoring service and rental service*. Unit on Evaluation, University of Illinois.

[4]The names and addresses of test publishers who issue catalogs are listed in the Appendix, page 12.

[5]Bloom, B. S. (Ed.) *Taxonomy of educational objectives; Handbook I: Cognitive domain*. New York: Longmans, Green and Co., 1956.

Readers may also be interested in a more recently published companion volume, which provides a classification of the affective goals of education and analyzes the relation of the affective to the cognitive domain: Krathwohl, D. R., Bloom, B. S., and Masia, B. B. *Taxonomy of educational objectives; Handbook II: Affective domain*. New York: David McKay, 1964.

further examination, still turn out to be lacking in essential qualities for our purposes and uses, but at least we have established the relevance of their content.

Next, we are ready to consider the relevance each test has for our purposes in testing achievement.

III. Purposes of Testing

Of course, one general purpose is to improve instruction. Although this is hopefully an ultimate result of our achievement testing programs, it is not very helpfully descriptive. Some of our intermediate purposes can be more clearly differentiated as: 1. *placement*, 2. *diagnosis*, 3. *assessment*, 4. *prediction*, and 5. *evaluation*.

1. We may often wish to *place* pupils in classes or sections according to their knowledge or skill in some subject area. We think of placement particularly in subject areas where there is a strong element of continuity, where later learnings seem to depend on earlier ones. In arithmetic, for example, it seems useless to try to teach multiplication to a pupil who has not yet learned how to add, to try to teach long division to a pupil who cannot do short division, and so on. Placement in this sense requires (a) recognition of the pupil's status in a particular content sequence and (b) provision of instruction that will start "where he is." Placement may be formalized in some schools by selective grouping of pupils into (for instance) separate arithmetic classes. In other schools, placement may consist of a single teacher's knowledge of her pupils' various positions in arithmetic achievement so that she can individualize instruction for each pupil or small group of pupils within her own class. Thus, placement may include placing the pupils in classes set up for various levels of instruction or placing instruction flexibly at levels suitable for different pupils. Therefore, for either type of provision, we shall use test scores to place or classify individual pupils in relation to each other and in relation to our curriculum content.

2. Or we may wish to *diagnose* pupils' weaknesses—gaps in their knowledge, understandings, and skills. This diagnosis will presumably help us to pinpoint individual remedial work and/or to revise our methods of teaching some elements that we deem important in the curriculum.

Whereas placement testing gives information of the same general nature as teachers' marks and ratings, diagnostic testing supplies a distinctly different kind of data. Diagnosis shifts the emphasis from outcomes to causes. Placement is concerned with the educational status of an individual; diagnosis is an effort to probe into and analyze specific deficiencies that may make for low status. Placement testing is generally appropriate for entire groups of pupils to ascertain in what segment of a range of achievement each pupil may belong. Diagnostic testing, on the other hand, is generally limited to those pupils who fall into one of the lower segments. Therefore, a placement test may be wide-ranging; but a diagnostic test is likely to be intensive rather than extensive, focused ("peaked") at a relatively low level. A placement test score furnishes inter-individual comparisons; a diagnostic test score furnishes intra-individual comparisons. For in diagnosis we are not so much concerned with an individual's performance in relation to the performance of other individuals; rather, we are concerned with his performance on what may be regarded as one function of some skill (e.g., reading) in relation to his performance on other functions pertaining to that skill. Thus, diagnostic tests tend to be long: a diagnostic reading test may require nearly as many items to analyze weakness in each reading function of a pupil as an entire placement test might require to designate his over-all reading status.

3. Or we may wish to *assess* effectiveness of teaching methods or other elements of the instructional program. Assessment of effectiveness is most likely to be concerned with pupil progress—growth, development, or change. For example, we may want to compare changes in reading comprehension of pupils who take Latin and pupils who do not; changes in arithmetic computation skills of pupils who take algebra and pupils who take general mathematics during an intervening year; changes in vocabulary of pupils who are drilled on word lists and pupils who are given wide reading experience instead; changes in history knowledge of pupils who have supervised study and pupils who study independently, of pupils who are taught by television and pupils who are taught in the flesh, of pupils who are grouped selectively according to some ability measure and pupils who are grouped heterogeneously; and so on.

If we put the three purposes (placement, diagnosis, assessment) on a time line, placement testing ordinarily comes first to determine the pupil's status or level—to find out what other pupils he should be grouped with. Diagnostic testing for the least able pupils represents a second stage in individualizing instruction—to break down educational deficiencies into components and thus to focus remediation where it can be expected to produce the greatest results. Then, to get some notion of the progress of all these pupils after a period of instruction (say, one year) comes an assessment test. Here the concern is not how the pupil (or group) stands in relation to other pupils (or groups) or how some of the pupils' (or group's) functions compare with other functions. Here we first want to compare the pupil (or group) *after* instruction with himself (or itself) *before* instruction. (Incidentally, in most school settings, there are some good arguments for focusing on group change rather than on individual change.) Then we may want to compare the amounts of change between different pupils or groups. As a practical matter, it is likely that we shall try to make one test do for both placement and assessment if we are interested in both purposes—i.e., the test that assesses progress during the preceding year (or two years, or whatever the interval may be) will often also serve to place pupils for their instruction during the succeeding year (or other period of time).

4. Or we may wish to *predict* a pupil's achievement or chances for success in each course that he may elect for the following year. Prediction is akin to placement in that it seeks, from a measure of present performance or status, a clue to future achievement. In fact, placement may be regarded as a special case of prediction—a case where there is a clear-cut dependence of future learnings on previous ones. But in most other cases of prediction from achievement test scores, *a priori* judgments must be carefully checked. In placement we are generally trying to specify differences in level of performance *within* a given subject area—e.g., should this pupil go into a faster or slower group in English? In prediction we may also seek to anticipate differences in level of performance *between* different subject areas (particularly when one or the other is to be elected)—e.g., is this student likely to do better

work in American history or in physics, in plane geometry or in French? We may choose to make assumptions of a relationship between, on the one hand, achievement test scores in social studies, general science, elementary algebra, and reading and—on the other hand—later success in American history, physics, plane geometry, and French, respectively. However, we had better not put too much mileage on such assumptions without checking them against empirical evidence. Of course even with such evidence we can't accept reported statistical relationships blindly but must also seek a theoretical explanation. The fact that a certain 8th-grade spelling test effectively predicted 10th-grade biology marks in one high school may be worth examining as a possibility for more general application, but there is no assurance that it would predict biology marks equally well in another school (where a biology teacher might be less devoted to the correct spelling of chlorophyll, cotyledon, and paramecium). Evidence of predictive validity of any test will look most promising when the reported studies have been based on a pupil population and a school curriculum which seem to resemble our own. In addition, if we are selecting an array of tests for purposes of differential prediction, we shall often be concerned with the appropriateness of comparisons between an individual's performances on tests taken about the same time in different subject areas.

5. Or we may say that we wish to *evaluate* our school. Evaluation is probably the purpose in achievement testing most frequently cited by administrators. For example, they may want to know—and perhaps be able to tell the board of education and the community—how good a job their schools are doing. Or they may want to study performance trends of their pupil population in certain achievement areas.

Evaluation of this sort is distinct from the evaluation implied in the other four purposes described above. It is true that we evaluate pupil achievement for placement, diagnosis, assessment, or prediction. Certainly each of these purposes involves some kind of judgment about high or low, good or bad, strong or weak. However, when we say that our purpose is to place, diagnose, assess, or predict, we are naming instructional actions that we shall take or educational decisions that we (or our pupils) shall make. We look to achievement tests to give us information that will provide a rational basis for these actions and decisions. This information is expressed in the form of test scores, plus statistics associated with test scores. All of these data are properly internal to our own school or school system. Thus, for placement, we are concerned with the achievement status of our pupils in relation to each other; for assessment, with the relation between end-of-course achievement and beginning-of-course achievement. For the purpose which might be labelled administrative evaluation, however, we are concerned with the status of our school in relation to other schools.

This purpose accounts for much of the interest in "national" norms. Somehow or other, the mid-point of a "national" norm has been taken to represent an appropriate standard of achievement. It is something a school can boast of being up to or better than. "Our 4th-grade pupils are well above the national average in reading and arithmetic" apparently has a good, praiseworthy sound. (To be below the "national average" seems less noteworthy—one rarely hears of such schools, although they must, by definition, be about equally numerous.) Although it seems a shame to deprive schools of this comfort, a comparison with so-called national norms is seldom very meaningful for judging a school. More to the point would usually be a comparison with a number of schools that may be assumed to draw from similar populations (examples: the independent secondary school norms compiled by the Educational Records Bureau; score distributions of different types of colleges classified in College Board Scores #2; provision for a college to select its own comparison group in the GRE Institutional Testing Program; statewide norms for schools participating in a state testing program; county norms for a cluster of rural schools; local norms for an urban school system). For evaluation, then, we must underline the importance of accurately characterizing our own school as discussed in I. SCHOOL POPULATION.

Now it seems clear that each of these five purposes involves, primarily, a comparison of some kind. Some tests may be suitable for only one of these purposes. Others may be suitable for more than one. It is worth noting here that, when we use a test for different purposes, we are not merely taking different kinds of action or making different kinds of decisions on the basis of a given test score or set of test scores. We are in fact using different kinds of standards, making different sorts of comparisons. Therefore, we shall seek different characteristics not only in the content of the tests but in the test scores themselves and in the accompanying statistical data. Thus, the bearing of our purpose on achievement test selection can perhaps best be clarified as we note the ways in which we *use* achievement test scores for these various purposes.

IV. Use of Test Scores

As we have indicated, one thing we do with test scores is compare them. We have already suggested five of these different kinds of comparisons. Let us summarize them here:

1. We may compare a pupil's achievement test score with the scores of other pupils in his school (as we noted under *placement*).

2. We may compare a pupil's (or group's) score on a subtest with his (its) score on another subtest (as in *diagnosis*).

3. We may compare a pupil's (or group's) achievement test score with his (its) score on a previous testing (as in *assessment*).

4. We may compare a pupil's score on one achievement test with his score on another achievement test (as in *prediction*).

5. We may compare a school's achievement test scores with achievement test scores of a comparable group of schools (as in *evaluation*).

And let us add these:

6. We may compare a pupil's (or group's) achievement test score with some fixed standard of mastery.

---

[6]Dyer, H. S., and King, R. G. *College board scores No. 2*. New York: College Entrance Examination Board, 1955. Fishman, J. *1957 Supplement to college board scores No. 2*.

7. We may compare a pupil's (or group's) achievement test score with his (its) scores on other kinds of tests (e.g., aptitude test, interest, personality, or attitude scales).

8. We may compare a pupil's achievement test score with other data about his achievement (e.g., school marks, ratings).

Let us assume that we have already screened tests in terms of our pupil population and our curriculum content and objectives. What test characteristics shall we then seek for each purpose and use we have listed?

1. Placement—comparison of pupil's achievement test score with achievement test scores of other pupils in his school.

For any test involving comparison among individuals, reliability must be, as we have seen, a prime consideration. How high should reliability be? This is a little bit like asking how tall should a basketball player be. Some textbooks will assert dogmatically that reliability coefficients should be "at least .90 and preferably around .95" for individual comparisons. Of course, as we have pointed out, no matter how high a coefficient, it may not be satisfactory if the reliability sample is not comparable with the population we want to test. *Granting such comparability*, a coefficient of internal consistency on the order of .9 seems, as a rule of thumb, reasonable to expect of a modern achievement test used for placement purposes. However, much depends on the subject area and grade level being tested and on the refinement of differentiation needed. Much depends also on how crucial the decision is. And much depends on what other data are available and how reliable they are. For example, if a forty-minute test has a reliability coefficient of .89, we should not ordinarily replace it with a 75-minute test for the sake of a reliability coefficient of .91.

The critical question concerning reliability is: how much difference between obtained scores of individuals can be taken to represent a difference in the same direction between their true scores? We know that an obtained score on a test includes a certain amount of error (due, for example, to chance factors). Therefore, two individuals scoring within a certain range of each other might, on another administration of the test or on an administration of an equivalent form, be more or less likely to change their rank in regard to each other. The Standard Error of Measurement indicates the likelihood of such changes for various score positions and differences on a test.

Now let us say that we want to divide one hundred pupils of the same grade into three arithmetic classes, intending that each class will proceed at a different pace. On almost any test of suitable content we may expect to find the highest-scoring and the lowest-scoring pupils clearly suitable for the fastest and slowest classes, respectively. But as we approach the middle scores from either extreme, we find it more difficult to make confident discriminations. For instance, there may be fifty pupils whose obtained scores are close enough so that we cannot feel comfortable in assuming that their true scores would place them in the upper, middle, or lower third. Thus, the 30th pupil who goes into the top group might, as far as the reliability of most tests is concerned, just as well belong in the middle or possibly even the low group.

A number of other groupings would also be possible. Of course, in most schools, tests will not determine for us the number of classes we may have. The budget is more likely to do that. But no matter how we divide our 100 pupils, we shall still inevitably have to reckon with a number of "borderline" cases. We may then want to ask, at which of the demarcation points are "mistakes" most damaging? If we decide that distinguishing the highest group from the others is crucial, we shall seek a test in which all the items tend to be quite difficult. If we are more concerned about identifying the lowest group, we shall look for a test in which all the items tend to be relatively easy. In short, the ideal placement test will be "peaked" so that all items tend to be of medium difficulty for students who are nearest the point where we want maximum discrimination. (If we want to discriminate at several points, we can expect that a test with items of middle difficulty for the entire group will give us maximum discrimination throughout the range.) Thus, data furnished by some publishers about item difficulties (usually in terms of "national" norms groups) give us a clue to the suitability of a placement test for our use.

Of course, not even the most carefully chosen test will eliminate all "mistakes" in classification. Therefore, we shall want to know how much confidence we can put in obtained score differences. So—having narrowed the field down to tests that have suitable content, apparently sufficient reliability, feasible length, and appropriate difficulty—we may prefer a test that shows us in a clear and helpful way how to reckon with possible measurement errors in score differences of various sizes. Thus, some tests facilitate proper comparisons between pupils' scores by reporting scores in a way that allows for the Standard Error of Measurement in various parts of the scale. (Example: converting each raw score to a percentile band or range instead of an exact percentile point.) Such score report forms and accessory interpretive aids help us to focus on obtained score differences of such a size that we can be reasonably sure (say, for instance, correct two-thirds of the time) that they represent real differences. In other words, we shall probably want to use for individual score comparisons a test that helps us to recognize its own limitations as well as its legitimate uses in differentiating between individual pupils.

2. Diagnosis—comparison of a pupil's score on a subtest with his score on another subtest.

A single test is rarely suitable both for survey purposes (as placement, assessment, prediction, evaluation) and for diagnostic purposes. Two intervening factors have already been suggested: length and difficulty level. It is true that an appropriate survey test might touch on each of the subtest areas in which we are interested for diagnosis. But even though the total score on a 30-40 minute test in some achievement area may be highly reliable, its subtests of (let us say) 3-8 minutes are comprised of relatively few items and are each almost always much less reliable. Therefore, in choosing a diagnostic test we must examine the reliability evidence for *each subtest*, not just the test as a whole. Obviously, then, we shall as a rule expect reliable diagnostic tests to be longer than survey tests. For example, one diagnostic reading test for 3rd-graders requires 3 hours, whereas 30 minutes is typical of survey reading tests. As for difficulty level, diagnostic testing is generally confined to a low achievement group. Therefore, a diagnostic test will usually be "peaked" lower (i.e., have many less difficult items) than a survey test. This fact also has a bearing on reliability: we shall want to know whether

the reliability sample was composed of pupils with low scores on a survey test. (See items 3 and 4, page 3.)

In selecting a diagnostic test, we shall be particularly interested in evidence of construct validity. We shall ask how the subtest areas have been defined. Is there reason to believe that learning in the general skill or subject area is directly affected by each function which has been discriminated and measured in a separate subtest? In other words, does a low score on a given subtest represent a truly distinct cause of low achievement in the entire area? If so, we shall usually want to find low correlations among the subtests themselves. Otherwise, if they are each measuring approximately the same thing, we can gain little more information from them than from a single-score survey test. Or, in case of substantial inter-correlations, is there evidence that the subtests represent successive stages, each dependent on the one preceding and each necessary for the one following? Thus, diagnosis would consist of determining in what stage the breakdown may have occurred. If the first stage malfunctions, there is no need or opportunity to observe the second and third stages. For example, if a poor reader can't discriminate between the letters **b** and **d**, **a** and **e**, **t** and **f**, he will obviously not be able to distinguish words like **bad** from **dab**, **rat** from **act**, **fad** from **tab**. But if we have satisfied ourselves that his letter recognition is satisfactory, we may then go on to check (among other possibilities) his knowledge of letter-sounds.

If we want to make comparisons among intra-individual sub-scores, it will be helpful if all of these sub-scores have about equally high reliability and are reported on the same score scale. But if each subtest applies to a successive stage, we can interpret each sub-score discretely: we can then accept greater variations in reliability and can use scores reported in different ways.

3. Assessment—comparison of score "after" with score "before."

Some score conversions frequently used in achievement tests—percentiles, age-equivalents, and grade-equivalents—generally should be handled very tentatively in comparisons involving amount of "growth," "change," or "progress." For one thing, the units of measurement are not necessarily equal at different parts of the scale. Thus a change of 5 percentile points from 50th to 55th percentile rank cannot be considered equal to a change of 5 percentile points from 90th to 95th percentile rank (the latter probably represents a greater increment of growth). Similarly, age or grade score units can be regarded as comparable only if we assume that for a given subject area the average progress made in any one year is equivalent to the average progress made in any other year. This assumption is often unlikely from the very fact that some subject areas receive more intensive teaching during one grade than during another. And, during a year, some skills appears to develop very dramatically at one age but to change almost imperceptibly at another. Also, above a certain grade the general ability level of the norms group is likely to change from one year to the next. Furthermore, we should recognize that the "month-units" used in grade- or age-equivalents are interpolations—based on another unlikely assumption: that progress in achievement is equally paced over the months of the school- or calendar-year.

Some tests do provide for converting scores into units designed to represent equal increments of achievement at different parts of the scale. "T" scores, "scale" scores, and "K" scores are examples of such efforts to define equal units.[7] Also, some tests provide a continuous scale spanning a wide range of grade levels. Such a scale may be especially useful in programs of testing spaced at different points over a number of grades. (Age- and grade-equivalents may appear to be continuous and wide-spanning but are not. For example, scores near the upper and lower extremes for pupils at a given age or grade are extrapolations and are generally not comparable from one level of the test to the next.) For use in assessment, then, it is likely that the assumptions underlying "T" scores, "scale" scores, and "K" scores are more acceptable than those underlying percentiles, age-equivalents or grade-equivalents.

Test users should consider also another problem in comparing test scores "before" and "after" a certain interval. If these scores were perfectly reliable the simple difference between a pupil's "before" and "after" scores could be taken to represent his gain during the interval. Since they are not perfectly reliable, some changes in individual pupils' scores would occur if the two testings were held without any interval. For example, some of the high scoring pupils would score lower and some of the low scoring pupils would score higher even though there was no substantial difference in the over-all score distribution of (let us say) the entire class tested. Since these changes (sometimes known as "regression effect") in individual scores would need be attributed only to the lack of perfect reliability in the test, it would be improper to say that these individuals had either gained or lost. Instead, the best estimate for any member of the class would be that he had neither gained nor lost regardless of any obtained difference (attributable to random errors of measurement) between his two scores. In practice, then, as Lord points out,[8] the best estimate of true gain for an individual student during an interval between testings will depend partly on the obtained difference between his scores and partly on the extent to which differences in the distributions of class scores exceed differences that would have been expected if the two testings had been administered at about the same time. Lord presents a formula and also a graphic procedure useful in estimating an individual's "true gain" by taking into account these two types of data. In general, since the differences between scores tend to be much less reliable than the scores themselves, many school users of achievement tests may be well advised to focus on assessment of change in average scores of a class (or other coherent group) rather than change in scores of individual pupils.

Incidentally, in assessment over a one-year span, it is generally possible and desirable to reuse for the second testing the same test form used for the first testing, thus eliminating error due to less than perfect equivalence of forms. If, however, an alternate form seems desirable so that teachers won't be tempted to drill their classes on specific test items, the effects of such error can be reduced by administering the forms in the order A B to one random half and in the order B A to the other random half of each group. Despite this precaution, needless to say, a very high coefficient of equivalence between forms will still be desirable.

---

[7]See Tiedeman, D. V. Has he grown? *Occupations*, November 1951.

[8]Lord, F. The measurement of growth. *Educational and Psychological Measurement*, 1956, 16, 421-37.

4. Prediction—comparison of a pupil's score on one achievement test with his score on another achievement test.

In prediction we are generally concerned with choices and decisions among various alternatives. Sometimes a single achievement test may be particularly appropriate for predicting a pupil's chances of success in a certain school subject. For example, in most schools, a spelling test is likely to offer the best single prediction of success in stenography. Perhaps in some schools an arithmetic computation test may be the most effective [illegible] for success for bookkeeping; a math test for [illegible]. In such situations, each test is regarded discretely; scores need not be compared with each other. [illegible] promising apparent relationship of the sort [illegible] can be studied and experience tables drawn up. From those tables a girl can be told that on the strength of her spelling test score, she probably has about 8 chances out of 10 (or 3 chances out of 10 or whatever it may be) of passing stenography. Even in the absence of high correlation between a test score and success in a certain subject area we can sometimes find evidence of valid cut-off points for prediction. Thus, a score above a certain point may be regarded as a necessary but not sufficient condition for passing a certain subject. Predictions of this sort require no direct comparisons among scores on achievement tests in different subject areas.

However, schools are often inclined to use achievement test "profiles" for prediction. A certain profile or pattern is supposed to represent more favorable chances for an individual in one alternative than in another. Such relationships are often fundamentally speculative or theoretical. Obviously, here again, a school will have to validate them empirically *after* using the tests. Otherwise, an attempt to predict from achievement test profiles by fiat involves numerous far-reaching assumptions, some of which have already been suggested (e.g., the applicability of reported validity studies to the population and setting in which the results will be used). However, before deciding whether various arrays of tests are even worth trying out for validation of pupils' profiles against their later performance, certain problems can be foreseen and avoided.

First of all, profiles tend to encourage direct manipulations and comparisons of scores on different tests. We are usually concerned with the "height" of a profile and may want to combine scores into a kind of "sight average" representing "total achievement" or achievement on some "meaningful cluster" (and then suggest to a pupil his chances of success in different curricula—say, college preparatory as opposed to commercial). We are also concerned with the "shape" of a profile so that we can say, for example, a pupil is higher in science than he is in social studies and is somewhere in between on arithmetic and English (and then perhaps suggest to him that he will probably do better in biology than in history). Let us for the moment disregard the (parenthetical) predictions we may make from combinations of scores or apparent differences between scores—these will be subject to specific validation after we use the tests. A prior question is: Are we justified in combining or differentiating between these scores at all? Are the scores truly comparable? To be so, an obvious first requirement is that all of the test results be expressed in the same kind of terms—e.g., all grade equivalents, all percentile ranks, etc. Then, we need to know whether the scores on all the tests refer to the same norms; we can not always assume that the 80th percentile on one norm is, in any meaningful way, higher than the 60th percentile on another norm. Next, we need to note that use of grade equivalent (or age-equivalent) scores for inter-test comparisons will very likely lead us astray. For example, we might be tempted to conclude that a 6th-grade pupil with grade-equivalent scores of 10.0 for reading and 8.5 for arithmetic is higher in reading than he is in arithmetic. Actually, he might rank at the 95th percentile of a norms group of 6th graders on both tests. (Grade-equivalent scores for arithmetic usually have a narrower range than grade-equivalent scores for reading because: 1. the content of arithmetic instruction tends to be more rigidly limited at each grade, and 2. a pupil's arithmetic performance is more likely to reflect classroom instruction.) Thus, for direct comparisons among achievement tests, we shall prefer percentile ranks, standard scores, or scaled scores to grade- or age-equivalents.

If the scores may be regarded as truly comparable we may then want to ask, How trustworthy are the differences between scores? Remember that a score difference is less reliable than a single score. If only one of the tests in a battery is less reliable than the others, comparisons and profiles involving that test are likely to have very limited usefulness for individual prediction. A profile that must be regarded as more "elastic" at some points than at others is very difficult to interpret, except in extreme cases. So, as a practical matter in selecting tests for "profile" predictions, we may again look with special favor on tests that provide built-in safeguards against over-refined discriminations. Then, regardless of the predictions we may ultimately base on score differences, at least we shall know when—within reasonable limits of confidence—a real difference can be taken to exist.

5. Evaluation—comparison of school achievement with achievement of other schools.

For evaluation, probably no test characteristic is more important than content. As we emphasized previously, if a test does not cover the learnings we deem most important it will have little usefulness to us for evaluation. We also stressed in an earlier section that meaningful comparisons require appropriate norms (pages 3–4, 7).

One other important point concerning norms is that comparisons of an average school score should be made with *school norms* and not with individual norms. School averages do not vary so widely as individual pupils' scores. Therefore, a school administrator will be in error if he tries to evaluate his school by comparing the average reading score of his 6th-graders with the 6th-grade norms describing the distribution of reading scores among individual pupils. For example, suppose that the average score of his 6th-graders coincides with the score ranked at the 75th percentile in the distribution of reading scores for a norms group of 7,000 6th-grade pupils taken from 200 schools. If the scores of these 6th-grade pupils from each school were averaged and the distribution of the 200 average scores were prepared, he might find that the mean score of his 6th-graders coincided with the 95th percentile of this distribution of mean school scores. So, in choosing an achievement test for school evaluation, we should generally look for one that provides *school norms* as well as pupil norms. Some test publishers name the schools constituting this norms group. In a few instances, for either school or individual pupil evaluation, it is possible for a test user to

select a *defined population of schools* for his norms group or even to select from a list of schools those which he wishes to comprise a norms group.

Since the reliability of mean scores is higher than the reliability of individual scores on any test, tests used solely for school evaluation can be relatively short.

6. For any of the purposes listed—placement, diagnosis, assessment, prediction, evaluation—we may sometimes measure pupils against a fixed standard of mastery. A single standard of mastery is like a "go" gage, and the "score" it gives need be merely dichotomous—"go or no go," "pass or fail," "does or doesn't," "yes or no." The pupil has come up to this standard or he has not, and there is no concern with the extent by which he may exceed it or fall short of it.

In discussing content validity, we pointed out that a test task itself might sometimes be selected so that perfect mastery is the standard. More frequently a cut-off point somewhat lower than perfection is set as the standard. In either case the main question is whether the standard has true utility. Thus, a standard is justified for *placement* if pupils who fall below that standard truly cannot "go"—cannot keep up with a faster-paced section; for *diagnosis*, if remedial work focused on a below-standard score in a subtest area will truly help a pupil "go"—will break the log jam to notable improvement in the entire subject area; for *assessment*, if the standard of growth with which actual progress is compared is truly to be expected in a certain period of time; for *prediction*, if the cut-off point truly minimizes error in distinguishing between alternative courses in which a pupil will "go" or "not go"—will have greater or less chance of success; for *evaluation*, if the standard is truly acceptable as one which the school population as a whole should achieve.

However, for most educational situations we cannot abide by a single fixed standard of mastery. Instead, our standards tend to be normative. We require not just a "go" or "no go"; rather we want to know where this pupil (or class or school) stands in relation to a comparable population. Thus, for most test scores, not a fixed standard of mastery but "mankind" is the proper measure of man.

7. For each of the purposes listed, we may find it helpful to compare a pupil's or group's achievement test scores with scores on another type of test (e.g., aptitude, "intelligence," interest). In some quarters it is believed that an "intelligence" test is a measure of innate capacity and that a ratio between achievement test and "intelligence" test performance will produce a meaningful "achievement quotient" or "educational quotient" for a pupil. Fortunately, such beliefs have by now fallen into sufficient disrepute that to challenge them here would be to demolish a straw man.

Incidentally, it is probably worth noting that a reverse assumption may, in some instances, be more palatable: an achievement test or battery based directly on specific school experience which is common to all pupils tested may be a better measure of innate capacity to learn than is a more generalized group intelligence test, performance on which may reflect more strongly the differences in pupils' home environment.

Be that as it may, a school can still find comparisons between "intelligence" and achievement test scores helpful. For example, in assessment and evaluation, scores on "intelligence" tests (perhaps better called scholastic aptitude or school ability tests) furnish concise descriptions of the general scholastic ability of a population; in placement, diagnosis, and prediction, aptitude and achievement test scores are often considered side by side on a pupil's cumulative record form. Unfortunately, such alignment may mislead a school person into making direct comparisons between aptitude and achievement scores that are not directly comparable even though expressed in similar terms such as percentile ranks. If the norms group for the aptitude test is not the same as the norms group for the achievement test, only a gross kind of comparison should be attempted (e.g., this pupil is in the top quarter of a certain population in verbal aptitude or quantitative aptitude or IQ, but in the lowest quarter of another population in spelling or science or algebra). To allow for more precise comparisons between achievement and aptitude test scores some achievement tests have been normed on the *same population* as certain aptitude or "intelligence" tests. If we are interested in making such comparisons, we should note in the test manuals whether such provisions (or others) have been made to equate certain achievement and aptitude test scores.

Also, we shall be interested in reports of correlations between achievement tests and aptitude or "intelligence" tests. For we shall want to know whether the extra time and cost involved in administering a number of achievement tests actually gives us appreciably more information than the group aptitude or "intelligence" test we probably administer anyhow. For example, if scores on a certain social studies or science achievement test are very highly correlated with scores on a verbal aptitude or "intelligence" test we are already using, we don't need that achievement test in addition to the verbal aptitude or "intelligence" test.

As for interest scores, we must not be misled by similarity of nomenclature. An interest score labeled "scientific" does not necessarily pertain to the same subject content as an achievement test in science. If we wish to assume that it does, we do so at our own risk.

8. Sometimes an achievement test is selected as a check on teachers' marks. Its use is then presumably for evaluation of individual pupils. Teachers' marks are believed to be widely influenced by a "halo" effect. The pupil who is likeable, neat, courteous, cooperative, etc., often seems to get consistently higher marks in all subjects than one who is not. An achievement test is assumed to give a measure of performance independent of such characteristics and also independent of differences among teachers at the same grade. Thus, we may want to use achievement test scores to modify the inferences that might be drawn from marks alone. However, for placement or prediction, it may be wise to bear in mind that if the criterion is to be a teacher's marks for the coming year, these marks, too, may represent a halo effect. Thus, teachers' marks for the past year may well be a better predictor than achievement test scores, especially if the prediction covers a time when the pupil will remain in the same school. If the prediction applies to a pupil who is changing schools, a standardized achievement test score is likely to contribute more effectively to the prediction.

Various systems have been used to put school marks and achievement test scores on the same scale. Such systems generally assume that both types of observation (marks and test scores) are normative—i.e., each mark or score can be placed in a normal distribution or some segment of

normal distribution of marks and test scores. However, teachers' marks in many subject areas are often based—at least implicitly—on fixed standards of mastery. This basis is revealed in the very fact that marks are frequently given as percentages—thus a pupil who is marked 75 in a subject is judged by the teacher to have learned about 75% of the teacher's standard of perfect mastery for that course. Teacher-made tests are most frequently scored in a similar way: 100 equals perfect performance, all correct, on a standard of perfect mastery; anything less than 100 represents a raw or weighted percentage of correct responses, with various cut-off points demarcating A from B, B from C, . . . and pass from fail. These marks for any given class may or may not be normally distributed. To assume that they are and to compare them directly with some kind of standard score or percentile rank on an achievement test would be to court error.

Then, must only an achievement test which is scored on a standard of mastery be selected for such comparisons between marks and test scores? Or must normalized test scores be used only if the class is marked "on the normal curve"? No, because as a rule fine distinctions and comparisons involving marks and test scores should not be made anyhow. Usually, coarse groupings of pupils (e.g., thirds, quarters) according to rank order based on either type of observation will suffice. Therefore, for comparison between marks and test scores, it is not necessary to select an achievement test that uses a scoring system comparable to the teacher's marking system. However, for most other uses, as we have seen, scoring systems, along with other characteristics, are quite relevant to achievement test selection.

Summary

There is no single achievement test or test battery that will be "best" for all pupil populations, all curriculum objectives, all purposes, and all uses. Even tests universally recognized as "good" are not equally "good" for different school settings, situations, and circumstances.

The recommended procedure for achievement test selection then consists of three phases:

1. study of our own school characteristics and testing needs;

2. analysis of characteristics and capabilities of available tests;

3. matching (a) the population of norms groups, of reliability samples, and of validity studies for each test to our own pupil population.
(b) the content of each test to our own curriculum content and objectives.
(c) the validity evidence, reliability data, scoring system, and interpretive material for each test to our own testing purposes and prospective uses of test scores.

This pamphlet has tried to explain, clarify, and illustrate these procedures. It has assumed that we shall also cut our testing cloth to suit our budget, our time schedule, and our staff. It has made no mention of such self-evident matters as format, administrative ease, and supplementary services.

Instead, it has tried to go beyond the superficial do's and don't's of test selection in the hope of assisting school people to make more knowledgeable use, as well as competent selection, of achievement tests. It is aimed at encouraging widespread staff participation in all three phases of selection—in the belief that all the school people who use an achievement test should have a hand in selecting it.

Appendix *(U.S. publishers of standardized tests who issue test catalogs)*

AMERICAN GUIDANCE SERVICE, INC., Publishers' Building, Circle Pines, Minnesota 55014

THE BOBBS-MERRILL COMPANY, INC., 4300 West 62nd Street, Indianapolis, Indiana 46206

BRUCE, MARTIN M., 340 Oxford Road, New Rochelle, New York 10804

BUREAU OF EDUCATIONAL MEASUREMENTS, Kansas State Teachers College, Emporia, Kansas 66801

BUREAU OF EDUCATIONAL RESEARCH AND SERVICE, C-6 East Hall, The University of Iowa, Iowa City, Iowa 52240

CALIFORNIA TEST BUREAU—See: CTB/McGraw-Hill.

CENTER FOR PSYCHOLOGICAL SERVICE, 1835 Eye Street, N.W., Washington, D.C. 20006

COMMITTEE ON DIAGNOSTIC READING TESTS, INC., Mountain Home, North Carolina 28758

CONSULTING PSYCHOLOGISTS PRESS, INC., 577 College Avenue, Palo Alto, California 94306

COOPERATIVE TESTS AND SERVICES, Educational Testing Service, Princeton, New Jersey 08540

CTB/McGRAW-HILL, West of Mississippi (Main Office): Del Monte Research Park, Monterey, California 93940. East of Mississippi: Princeton Road S-2, Highstown, New Jersey 08520

EDUCATIONAL TEST BUREAU—See: American Guidance Service, Inc.

EDUCATORS' EMPLOYERS' TESTS AND SERVICES ASSOCIATES, 120 Delzel Place, Cincinnati, Ohio 45219

FOLLETT EDUCATIONAL CORPORATION, 1010 West Washington Boulevard, Chicago, Illinois 60607

GUIDANCE TESTING ASSOCIATES, 6516 Shirley Avenue, Austin, Texas 78752

HARCOURT BRACE JOVANOVICH, INC., 757 Third Avenue, New York, New York 10017

HIGH SCHOOL TESTING SERVICE, Measurement and Research Center, Purdue University, Lafayette, Indiana 47907

HOUGHTON MIFFLIN COMPANY, 110 Tremont Street, Boston, Massachusetts 02107

INSTITUTE FOR PERSONALITY AND ABILITY TESTING, 1602 Coronado Drive, Champaign, Illinois 61820

OHIO TESTING SERVICES (Ohio Scholarship Tests), Division of Guidance and Testing, State Department of Education, 751 Northwest Boulevard, Columbus, Ohio 43212

PERSONNEL PRESS, INC., 191 Spring Street, Lexington, Massachusetts 01273

PERSONNEL RESEARCH INSTITUTE, Case Western Reserve University, 11105 Euclid Avenue, Cleveland, Ohio 44106

THE PSYCHOLOGICAL CORPORATION, 304 East 45th Street, New York, New York 10017

PSYCHOMETRIC AFFILIATES, Box 3167, Munster, Indiana 46321

PUBLIC SCHOOL PUBLISHING COMPANY—See: Bobbs-Merrill Company, Inc.

SCHOLASTIC TESTING SERVICE, INC., 480 Meyer Road, Bensenville, Illinois 60106

SCIENCE RESEARCH ASSOCIATES, INC., 259 East Erie Street, Chicago, Illinois 60611

SHERIDAN PSYCHOLOGICAL SERVICES, INC., 10481 Santa Monica Boulevard, West Los Angeles, California

STOELTING COMPANY, 424 North Homan Avenue, Chicago, Illinois 60624

TEACHERS COLLEGE PRESS, Teachers College, Columbia University, 525 West 120th Street, New York, New York 10027

UNIVERSITY BOOKSTORE, Purdue University, 360 State Street, West Lafayette, Indiana 47906

WESTERN PSYCHOLOGICAL SERVICES, Box 775, Beverly Hills, California 90213

WORLD BOOK COMPANY—See: Harcourt Brace Jovanovich, Inc.

03047 • T63P25 • 278919